图书情报工作
第63卷 第18期 2019年9月

基于依存句法分析的中文专利候选术语选取研究*

■ 俞琰[1,2] 陈磊[1] 姜金德[3] 赵乃瑄[1]
[1] 南京工业大学信息服务部 南京 210009 [2] 东南大学成贤学院计算机工程系 南京 211816
[3] 南京晓庄学院商学院 南京 211171

摘要：[目的/意义]针对中文专利候选术语选取方法存在需要对不同的数据集分别制定不同的模式匹配规则、专利术语抽取准确性不高等问题，本文提出基于依存句法分析的中文专利术语选取方法，以提高中文专利术语抽取准确性。[方法/过程]主要包括依存句法分析、剪枝、生成依存子树等三个主要步骤。首先对中文专利进行依存句法分析，得到依存树，对依存树进行剪枝，去除不符合要求的依存关系，生成依存子树，从中选取连续词串作为候选术语，以抽取中文专利术语。[结果/结论]实验结果表明，与已有的中文专利候选术语选取方法相比，本文提出的基于依存句法分析的中文候选术语选取方法能够有效地提高中文专利术语抽取的准确性。

关键词：术语抽取 依存句法分析 中文候选术语选取
分类号：G202
DOI：10.13266/j.issn.0252－3116.2019.18.013

1 引言

专利文献蕴含着丰富的各个领域问题的解决方案，有效的专利文献分析能够判断领域技术热点、识别领域核心技术，预测领域技术发展趋势，帮助研发人员从中获得启发与借鉴，从而缩短创新设计时间、节约创新设计经费。其中，专利文献中的术语为专利文献分析提供了结构化知识单元，体现和承载了专利文献的技术信息，成为诸多专利文献分析的关键组成部分，因此，如何有效地从专利文献中自动抽取术语是专利分析中重要的研究内容。

目前，专利术语抽取方法通常包括候选术语选取与候选术语排序两个步骤，即首先从语料中选取候选术语；然后利用统计信息计算候选术语成为术语的可能性，按照可能性的高低进行排序，满足一定要求的候选术语则被认定为术语。其中，针对中文专利术语抽取任务，通常采用词性模式匹配方法选取候选术语，如“形容词＋名词”、“动词＋名词”等模式，以选取中文专利候选术语。但这种方法主要存在2个问题：①针对不同的中文专利文本集需要人工定义不同的匹配规则，实现难度较大；②在选取正确候选术语的同时也可能引入过多的非术语词串，例如，依照“动词＋名词”模式，虽然可以正确地选取“氧化 石墨烯”等候选术语，但是也引入“添加 粉末”等非术语词串。

依存句法分析通过语句单位内词语间的依存关系揭示词语间的语义修饰关系，进而实现对语义的理解，可以较为有效地弥补单纯依靠词性手段难以触及深层语义关系的不足。因此，本文首次将依存句法分析引入中文候选术语选取研究之中，提出基于依存句法分析的中文专利候选术语选取方法，以提高中文专利术语抽取的准确性。实验结果表明，与已有的中文专利候选术语选取方法相比，本文提出的基于依存句法分析的中文候选术语选取方法能够有效地提高中文专利术语抽取的准确性。

2 相关研究

2.1 术语抽取

术语指某一专业知识活动领域中一般（具体或抽

* 本文系教育部人文社会科学规划项目“大数据时代技能知识图谱构建研究”（项目编号：16YJAZH073）和国家社会科学基金一般规划项目“大数据时代支持创新设计的多维度多层次专利文本挖掘研究”（项目编号：17BTQ059）研究成果之一。
作者简介：俞琰（ORCID：0000－0002－9654－8614），副教授，博士，E-mail：yuyanyuyan2004@126.com；陈磊（ORCID：0000－0002－5504－7493），硕士研究生；姜金德（ORCID：0000－0002－5504－7493），教授，博士；赵乃瑄（ORCID：0000－0001－9072－7315），馆长，教授，博士。
收稿日期：2019－01－22 修回日期：2019－04－14 本文起止页码：109－118 本文责任编辑：杜杏叶

象)理论概念[1-2],是专业领域知识系统中的重要组成部分,传达了大量的领域知识。通过对术语的了解可把握一门专业领域技术的精髓所在。术语抽取是指从文本中自动发现术语的过程。

目前,术语抽取方法可分为无监督方法和有监督方法两大类。无监督方法通常利用语言学与统计学相结合的方法,从文本集中抽取术语,具有较少人工干预、较强的适用性和一致性等优点;有监督方法采用机器学习方法,如最大熵模型[3]、条件随机场[4-5]等,通过学习训练文本特征,以抽取术语。有监督方法能够弥补无监督方法无法识别低频术语的缺陷,方法抽取准确率和召回率较高,但需要大规模人工标注语料作为训练数据,对训练语料的规模与质量要求较高,并且,有监督方法还不成熟,需要进行更多的尝试与验证[6]。目前专利文献没有针对性的、完备的、大规模标注语料,且随着科技的快速发展,大量新术语不断涌现,无监督方法可以在极低人工干预下抽取术语,是克服有监督方法标注语料获取困难的有效途径之一。因此,本文着重研究使用无监督方法抽取专利术语。

无监督方法虽然很多,但通常遵循"先候选,再排序"的流程,即,候选术语选取和候选术语排序两个步骤。以下分别介绍这两个步骤。

2.1.1 候选术语选取 总的来说,候选术语选取方法可分为 *n*-gram 过滤(*n*-gram filtering)[7-9]、名词短语分块(NP-chunking)[1]和词性模式匹配(POS tag pattern)[10-13]等三种方法。

n-gram 过滤通常先去除停用词、语义信息较少的词(如助词、语气词等)或者人工选择构词能力较差的词,得到文本串片段,然后进行遍历得到所有 *n* 元连续词语序列,按照一定规则选出符合要求的多元词组,如保留词频高的词语。这种方法具有实现简单且灵活可设置多元词组长度等优点,但也存在非术语词串过多从而影响术语抽取准确率的问题。

因术语通常为名词性短语,故名词短语分块方法从词性标注后的文本序列中识别出名词短语。名词短语的词性规则通常遵循特定的排列模式,如"形容词+名词"模式。因此,通常采用模式匹配结合句法规则来识别名词短语。这种方法简单快速,但由于中文名词短语中修饰词的词性规则较为复杂,不仅仅局限于形容词和名词,因此,这种方法往往应用于英文术语抽取。

词性模式匹配的基本思想与名词短语分块相同,均默认术语的词性序列遵循特定排列模式。不同之处在于,词性模式匹配通过定义更加复杂的匹配模式。其优点是能够针对中文文本特点,指定有针对性的匹配规则,因此这种方法是目前中文候选术语选取的主流方法之一。但是该类方法存在需要针对不同中文数据集人工定义不同匹配规则、在选取正确候选术语的同时也可能引入过多非术语词串等问题。

2.1.2 候选术语排序 候选术语排序主要使用术语性(Termhood)和单元性(Unithood)度量候选术语成为术语的可能性。

术语性从术语的隶属度出发,衡量一个候选术语与特定领域的相关程度。常用的统计量有词频[14]和C-value[1]及其变体等方法。其中,词频方法根据候选术语在语料集中出现的频次度量候选术语的领域相关程度。但词频方法对低频术语的抽取没有引起足够的重视。C-value 方法为词频方法的改进,它考虑了短语的嵌套性,统计信息包括候选术语的词频、词长、包含当前候选术语的更长候选术语的频次和个数。C-value 方法简单、适用性强,具有语言和领域无关性。然而,C-value 方法仍然以候选术语频次为主要依据,不能有效地过滤一些高频非术语词串以及正确抽取低频术语。针对这些问题,一些研究尝试对 C-value 方法进行改进。如 PCC-value[15]将候选术语的文档频次融入 C-value 计算之中;STC-value[16]利用与候选术语具有相同术语部件的相似候选术语信息,以提高 C-value 术语抽取效果。

单元性度量候选术语结构的稳定程度,即候选术语内部各组成部分之间的结合强度。其中,互信息是一种常用的单元性指标。互信息通过计算候选术语中各词成分的共现频次来衡量这些成分之间的依赖程度[17]。互信息方法能够较好地反映字串之间的结合强度,但会过高估计低频且总是相邻出现的字串间的强度[18],一些研究提出 PMI^k[19]以及 EMI[20]等互信息变体方法,以改进互信息方法过高评估低频候选术语强度的问题。

2.2 依存句法分析

依存句法分析的基本假设是:句法结构本质上包含词对间的关联,一个词支配另一个词,这种支配与被支配的关系被称为依存关系。依存句法分析认为语句中核心动词是支配其他词语的中心成分,而它本身却不受其他任何词语的支配,所有受支配词语都以某种依存关系从属于支配词。依存句法分析通过分析语句中词语之间的依存关系揭示其句法结构,发现语句语法特征和语义联系。根据依存句法分析公理[21-23],在

一个完整的语句中,任何一个词语都不能依存于2个或2个以上的其他词语,所有的语义联系相互交织的结果将语句的线性结构层次化,构造成为一棵依存树,从而反映出句子中词语间的语义修饰关系,且与成分的物理位置无关。依存句法分析具有表示简洁、存储空间小、可计算性好、中心驱动、词性依赖较小、依存关系具有普遍性等特点,较适合灵活的中文词序[24]。依存句法分析本身没有规定要对依存关系进行分类,但为了丰富依存结构表达的句法信息,在实际应用中,一般会对依存树的边加上不同的标记。其中,哈尔滨工业大学信息检索研究室语言技术平台 LTP(Language Technology Platform)[25]的依存关系标注体系具有依存关系定义关系数量较少、易于理解等优点。

由于依存句法分析通过分析语句中词语间的依存关系揭示词语间的语义修饰关系,可以反映长距离的搭配信息并与词语的物理位置无关,因此依存句法分析被广泛应用于情感分析[26-28]、实体关系抽取[29-31]、自动问答[32-34]、触发词识别[35-37]等自然语言处理任务之中。例如,在情感分析相关研究中,邓淑卿等[28]提出基于句法依存规则以及词性特征的情感词识别模型,利用8种关系模式作为情感词匹配候选模板,以识别京东商城 iPhone 6s 手机评论中的情感词。在实体关系抽取相关研究中,李明耀等[30]针对中文语法错综复杂,表达方式灵活,语义多样等固有性质的限制,提出一种开放式中文实体关系抽取方法,通过依存句法分析的依存关系判断语句是否为动词谓语句。在自动问答相关研究中,刘雄等[33]将依存树中边上的依存关系标签改为表征问句分解信息的分解标签,进而生成子问句,以增强问答系统理解复合事实型问句的能力。在触发词识别的相关研究中,高源等[36]利用依存句法分析,提出触发词－实体描述对的方法,以提高触发词的抽取召回率。

综上所述,总的来说,术语抽取研究主要集中于候选术语排序算法改进中,对候选术语选取研究关注过少。特别是中文候选术语选取方法存在制定模式匹配规则困难、需要针对不同的数据集分别制定不同的模式匹配规则、专利术语抽取准确性不高等问题。依存句法分析通过语句内词语间的依存关系揭示词语间的语义修饰关系,进而实现对语义的理解,可以较为有效地弥补单纯依靠词性手段难以触及深层语义关系的不足,因此,本文提出一种基于依存句法分析的中文专利候选术语选取方法。

3 基于依存句法分析的中文专利候选术语选取方法

本文引入依存句法分析,形成基于依存句法分析的中文候选术语选取方法。方法主要包括依存句法分析(第3.1节)、剪枝(第3.2节)、生成依存子树(第3.3节)等三个主要步骤。首先对收集的中文专利文本集进行依存句法分析,得到依存树,对依存树进行剪枝,去除不符合要求的依存关系,生成依存子树,从中选取连续词串作为候选术语,得到可以进行候选术语排序的候选术语集,以抽取中文专利术语。

3.1 依存句法分析

依存句法分析通过分析语句中词语之间的依存关系,揭示语句的句法结构。其中,依存关系可以使用有向弧表示,由支配词指向其被支配词,并且依存句法分析认为语句中的支配者是核心动词。根据依存语法公理,在一个完整的语句中,依存句法分析将语句的线性结构层次化,构造成为依存树。本文据此给出依存树的定义。

定义1(依存树):依存树记为 $T=(V, A, R)$,其中 V 为结点集合,表示语句中词语;A 为有向弧集合,表示词语间依存关系,弧的出发端为依存关系的支配词,弧的指向端为依存关系的被支配词;R 为依存树根结点,为语句核心动词,T 满足:

①R 结点的入度为0;

②除 R 之外结点的入度为1;

③从 R 到任一结点有一条有向通路。

图1为使用哈尔滨工业大学语言技术平台发布的依存句法分析器,对语句“本发明主要用于制备四氧化铁负载氮掺杂石墨烯复合材料”与“本发明涉及一种共掺杂聚吡咯材料及其制备方法和应用”进行依存句法分析之后得到的依存树 T1 和 T2。其中,Root 分别指向两个语句的核心动词“用于”和“涉及”,弧上标注表明依存关系的类别,主要依存关系类别见表1。结点下的字母表示词性,本文主要使用的词性见表2。由图1可见,依存句法分析在分词的基础上给出了词语间关系以及语句的浅层句法结构,这为中文专利候选术语选取提供了依据。

3.2 剪枝

由于中文专利术语一般为名词短语,依存树中一些依存关系通常不会出现在名词短语之内,这些关系会引入大量噪声,影响候选术语选取的结果,因此,本文提出对依存树进行剪枝,目的是在选取候选术语之

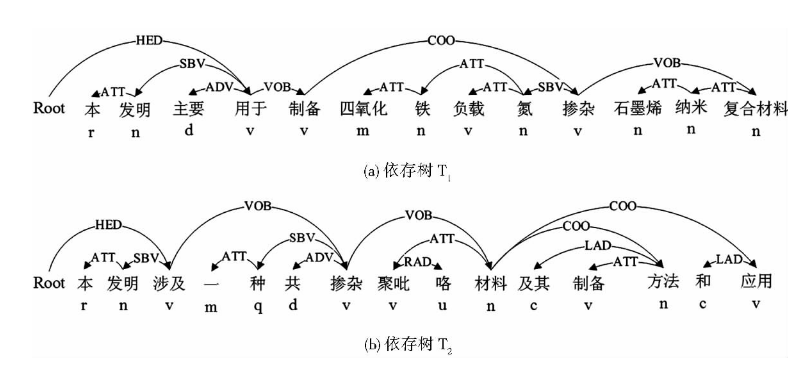


图 1　依存树示例

表 1　依存关系

关系类型	关系标记	样例
主谓关系	SBV	我送她一束花(我←送)
动宾关系	VOB	我送她一束花(送→花)
间宾关系	IOB	我送她一束花(送→她)
前置宾语	FOB	他什么书都读(书←读)
兼语	DBL	我请我吃饭(请→我)
定中关系	ATT	红苹果(红←苹果)
状中关系	ADV	非常美丽(非常←美丽)
动补结构	CMP	做完了作业(做→完)
并列关系	COO	大山和大海(大山→大海)
介宾关系	POB	在贸易区内(在→内)
左附加关系	LAD	大山和大海(和←大海)
右附加关系	RAD	孩子们(孩子→们)
核心关系	HER	我送她一束花(→送)

表 2　主要词性

词性	词性标记	词性	词性标记
名词	n	数字	m
动词	v	区分词	b
名动词	vn	连词	c
形容词	a	后缀	k
副词	d	量词	q
代词	r	助词	u

前尽量减少无用信息。对表 1 中的依存关系进行分析后发现,中文专利术语内词的关系主要为定中关系(ATT)、并列关系(COO)、左附加关系(LAD)和右附加关系(RAD)等四类依存关系。此外,中文专利术语一般由包含丰富领域信息的词语构成,通常不包含停用词,因此本文选取哈尔滨工业大学停用词表和人工选取的若干典型中文专利停用词,如,“发明”、“方法”等作为停用词。基于以上分析,本文提出两个剪枝规则对依存树进行剪枝:

剪枝规则 1:去除定中关系(ATT)、并列关系(COO)、左附加关系(LAD)和右附加关系(RAD)之外的其他依存关系;

剪枝规则 2:去除包含停用词的依存关系。

图 2 为对图 1 中的依存树 T1 和 T2 进行剪枝,其中灰色的弧度表示根据剪枝规则 1 和剪枝规则 2 去除的依存关系,灰色词语表示停用词。

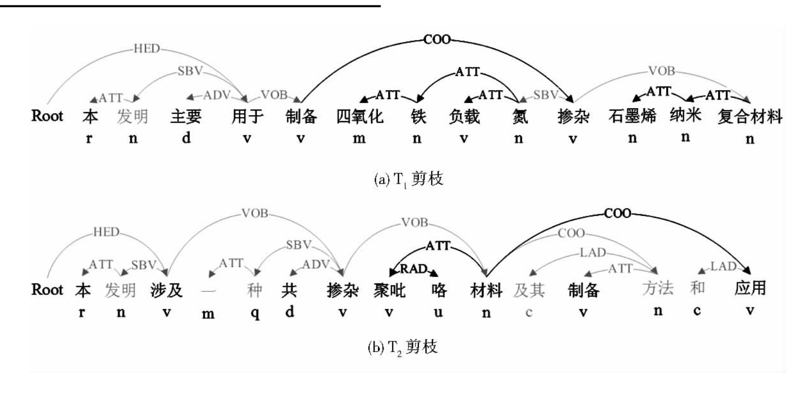


图 2　依存树剪枝

3.3　生成依存子树

由于中文术语通常为以名词或动词为核心,其他各种成分修饰的短语。因此,本文提出依存子树的概念,以选取中文候选术语,其定义如下:

定义 2(依存子树):给定一棵依存树 T = (V, A, R),则依存子树 T′ = (V′, A′, R′)满足如下条件:

① $V' \subseteq V$, $A' \subseteq A$, $R' \in V'$;

② R′结点的入度为 0;

③ 除 R′之外结点的入度为 1;

④ 从 R′到任一节点有一条有向通路;

俞琰,陈磊,姜金德,等.基于依存句法分析的中文专利候选术语选取研究[J].图书情报工作,2019,63(18):109 – 118.

⑤ R′为名词或动词等实词。

根据定义 2,对图 2 剪枝后的依存树 T_1 和 T_2 进行筛选,生成如图 3 所示的依存子树 $T_{1,1}$ ~ $T_{1,8}$ 和 $T_{2,1}$ ~ $T_{2,3}$。其中,由于后续欲比较的候选术语抽取方法均选取大于 1 个词的词串构成的短语,因此,本文仅选取 $|V|>1$ 的依存子树,即至少包含 2 个词。

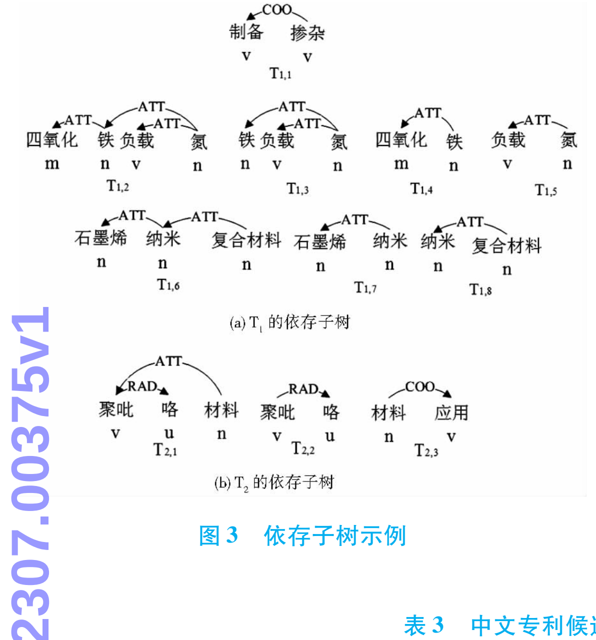


图 3　依存子树示例

根据生成的依存子树,选取连续词串作为中文专利候选术语。例如,图 3(a)的依存子树 T1,1 为非连续词串,因此舍弃,而剩余的依存子树 T1,2 ~ T1,8 均为连续词串,因此生成 7 个候选术语;图 3(b)中的依存子树 T2,1 和 T2,2 中的词串为连续词串,T2,3 的词串为不连续词串,所以舍弃,最终生成 2 个候选术语。

表 3 与表 4 给出了使用 n-gram[7]、名词短语分块[1]、词性模式匹配名词[15]和本文方法对语句“本发明主要用于制备四氧化铁负载氮掺杂石墨烯复合材料”和语句“本发明涉及一种共掺杂聚吡咯材料及其制备方法和应用”分析后得到的候选术语结果。由表 3 和表 4 可见,n-gram 方法虽然包含所有正确候选术语,但是存在错选候选术语过多的问题;名词短语分块方法则会造成漏选问题;词性模式匹配方法同时存在漏选且错误候选术语过多等问题;相对而言,本文提出的基于依存句法分析的方法能够找到较多正确的术语,且错误术语相对较少,为后续候选术语排序打下坚实基础。

表 3　中文专利候选术语选取方法比较示例 1

方法	n-gram(n = 2 ~ 4)	名词短语分块	词性模式匹配	基于依存句法分析方法
正确候选术语	四氧化 铁 纳米 复合材料 石墨烯 纳米 复合材料	纳米 复合材料 石墨烯 纳米 复合材料	纳米 复合材料 石墨烯 纳米 复合材料	四氧化 铁 纳米 复合材料 石墨烯 纳米 复合材料
错误候选术语	主要 用于 用于 制备 制备 四氧化 铁 负载 负载 氮 氮 掺杂 掺杂 石墨烯 石墨烯 纳米 主要 用于 制备 用于 制备 四氧化 制备 四氧化 铁 四氧化 铁 负载 铁 负载 氮 负载 氮 掺杂 氮 掺杂 石墨烯 掺杂 石墨烯 纳米 主要 用于 制备 四氧化 用于 制备 四氧化 铁 制备 四氧化 铁 负载 四氧化 铁 负载 氮 铁 负载 氮 掺杂 负载 氮 掺杂 石墨烯 氮 掺杂 石墨烯 纳米 掺杂 石墨烯 纳米 复合材料	石墨烯 纳米	石墨烯 纳米 铁 负载 氮 掺杂 负载 氮 掺杂 石墨烯 掺杂 石墨烯 纳米 铁 负载 氮 氮 掺杂 石墨烯 掺杂 石墨烯 纳米 复合材料 负载 氮 掺杂 石墨烯	四氧化 铁 负载 氮 铁 负载 氮 负载 氮 石墨烯 纳米
漏选候选术语	(无)	四氧化 铁	四氧化 铁	(无)

表 4 中文专利候选术语选取方法比较示例 2

方法	n-gram(n = 2 ~ 4)	名词短语分块	词性模式匹配	基于依存句法分析方法
正确候选术语	聚吡 咯	(无)	(无)	聚吡 咯
错误候选术语	共 掺杂 共 掺杂 聚吡 共 掺杂 聚吡 咯 共 掺杂 聚吡 咯 材料 掺杂 聚吡 掺杂 聚吡 咯 掺杂 聚吡 咯 材料 聚吡 咯 材料 咯 材料 制备 方法	(无)	制备 方法	聚吡 咯 材料
漏选候选术语	(无)	聚吡 咯	聚吡 咯	(无)

4 实验

4.1 数据集

为了验证提出模型的可行性与有效性,本文选取石墨烯专利文献进行实验。石墨烯是已知材料中最薄的一种,因其具有独特的结构,集优异的光学、化学、电学、力学等特征于一身,迅速成为物理学、化学和材料学等领域最热门的研究主题之一。石墨烯也被认定为新型潜力材料,具有可观的经济效益和广泛的产业化应用前景,可广泛应用于新型复合材料、储能装置、电极、超灵敏传感器、新型催化剂等领域。实验基于中国国家知识产权局专利数据库,以"石墨烯"关键词检索中国近 5 年来(2014 - 2018 年)的有效中国发明公开专利(检索日期为 2018 年 11 月 15 日),共获得 6 445 条有效中国发明公开专利,以其题名和摘要作为专利文本数据集。

4.2 评估标准

鉴于专利文本数据较多,采用准确率作为评估指标,即评估被抽取的前 N 条术语的正确性[38]:

$$准确率 = \frac{正确抽取的术语数}{抽取的术语数} \times 100\% \qquad 式(1)$$

本实验 N 分别取 200 - 2 000,采用人工方式对实验结果进行判断,为了避免主观性和领域知识的局限性,利用百度百科、维基、互动百科等知识网站,结合专家评估的办法判断被抽取术语的正确性。

4.3 实验结果

4.3.1 候选术语选取方法对术语抽取效果影响比较

实验首先使用典型的候选术语选取方法与本文提出的方法进行比较,欲比较的候选术语选取方法如下:

(1)*n*-gram[7]:先去除停用词、语义信息较少的词(如助词、语气词等)、人工选择构词能力较差的词(如,发明,方法等),得到文本串片段,然后进行遍历得到所有 n 元连续单词序,实验设定 $n = 2 - 6$。

(2)NP[1]:名词短语分块方法认为术语的词性规则通常遵循特定的排列模式,如"形容词 + 名词"、"名词 + 名词"模式。实验使用的正则表达式为 $(a|n)^{+}n$。

(3)pos1[15]:使用表 5 所示的词性模式匹配规则选取中文候选术语。

表 5 词性模式匹配规则[15]

长度	词性模式匹配规则
2 词	n + n、n + v、v + n、a + n、d + n、b + n
3 词	n + n + n、v + n + n、n + v + n、v + v + n、b + v + n、n + m + n
4 词	n + n + n + n、n + n + v + n、v + n + n + n、v + n + v + n、n + v + v + n、v + v + n + n、v + n + b + n
5 词	v + v + n + n + n、d + v + n + n + n、m + v + m + n + n、b + v + n + v + n、n + n + v + n + n、a + n + v + n + n
6 词	n + n + c + vn + n + n、n + n + vn + c + vn + n、n + n + u + b + vn + n、vn + n + vn + c + vn + n、l + vn + k + n + vn + n、n + vn + u + n + vn + n

(4)pos2[10]:使用表 6 所示的词性模式匹配规则选取中文候选术语。

表 6 词性模式匹配规则[10]

长度	词性模式匹配规则
2 词	n + n、v + n
3 词	n + n + n、n + b + n
4 词	b + m + n + n、b + n + n + n
5 词	n + n + n + vn + n、m + n + b + vn + n
6 词	b + n + b + n + n + n、n + n + u + b + vn + n

(5)dep:本文第 3 节提出的基于依存句法分析的候选术语选取方法。

为了进行评估,实验使用常见的两种候选术语排序算法 C-value 和 PMI 进行候选术语排序。总的,将欲使用的中文专利术语抽取方法分为两组,如表 7 所示:

表 7 欲比较中文专利术语抽取方法

组号	候选术语选取方法	候选术语排序方法	术语抽取方法表示
第一组	n-gram NP pos1 pos2 dep	C-value	n-gram + C-value NP + C-value pos1 + C-value pos2 + C-value dep + C-value
第二组	n-gram NP pos1 pos2 dep	PMI	n-gram + PMI NP + PMI pos1 + PMI pos2 + PMI dep + PMI

实验结果如图 4 和 5 所示。图 4 为第一组基于 C-value 候选术语排序的不同候选术语选取方法比较结果。由图 4 可见,在 5 种候选术语选取方法中,*n*-gram

+C-value 的准确率、召回率和 F 值均最低，例如，在 N =1 000 时，其准确率仅为 35.75%；NP + C-value 方法的准确率较 n-gram + C-value 有所提升，如在 N = 1 000 时，其准确率较 n-gram + C-value 方法提升了 11.03%，但低于 pos1 和 pos2 两种方法；pos1 + C-value 和 pos2 + C-value 方法的准确率类似，pos2 + C-value 优于 pos1 + C-value，pos2 + C-value 方法在 N = 1 000 时，其准确率较 *n*-gram + C-value 方法分别提升了 13.60% 和 17.58%；本文提出的基于依存句法分析的方法获得了最高的准确率，如在 N = 1 000 时，准确率比 *n*-gram + C-value 提高了 24.37%。在第二组以 PMI（图 5）作为候选术语排序方法的比较中，得到了类似的结果，*n*-gram 方法准确率最低，NP 次之，pos1 与 pos2 优于 NP，dep 方法获得最好的准确率。由两组实验结果可以看出，本文提出的基于依存句法分析的中文候选术语选取方法明显优于传统的候选术语选取方法，能够获得更高的中文专利术语抽取准确率。

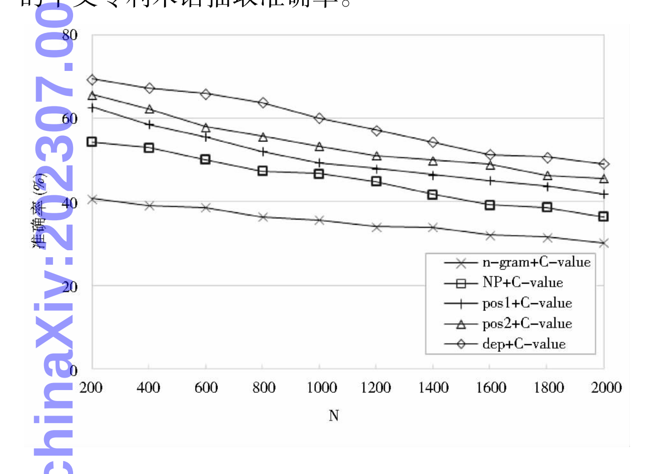


图 4 第一组基于 C-value 的不同候选术语选取方法准确率比较

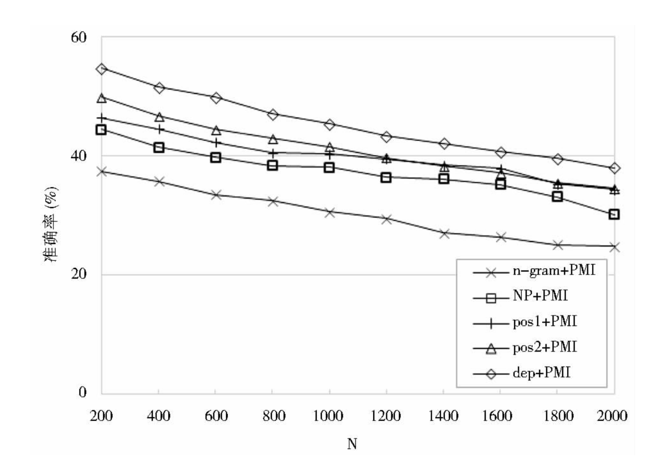


图 5 第二组基于 PMI 的不同候选术语选取方法准确率比较

4.3.2 候选术语选取与候选术语排序对术语抽取效果影响比较 无监督术语抽取方法通常包括候选术语选取和候选术语排序两个步骤，目前的相关研究主要集中于候选术语排序算法改进上。本部分尝试探究候选术语选取与候选术语排序对最终中文专利术语抽取准确率的影响程度。

C-value 和 PMI 是两类较为典型的候选术语排序度量方法，有较多的研究与改进方法。因此，本文采用传统候选术语选取方法中效果较好的 pos2 方法（即，使用表 5 所示的词性匹配规则方法选取候选术语），使用 pos2 和 C-value 作为第一组术语抽取的基准方法，使用 pos2 和 PMI 作为第二组术语抽取的基准方法。比较分别改进候选术语选取方法和改进候选术语排序方法对中文专利术语抽取准确率的影响程度。具体欲比较的方法如表 8 所示：

表 8 欲比较的改进候选术语选取方法与候选术语排序方法

组号	方法类型	候选术语选取方法	候选术语排序方法	术语抽取方法表示
第一组	基准方法	pos2	C-value	pos2 + C-value
	改进候选术语排序方法	pos2	PCC-value	pos2 + PCC-value
		pos2	STC-value	pos2 + STC-value
	改进候选术语选取方法	dep	C-value	dep + C-value
第二组	基准方法	pos2	PMI	pos2 + PMI
	改进候选术语排序方法	pos2	PMIk	pos2 + PMIk
		pos2	EMI	pos2 + EMI
	改进候选术语选取方法	dep	PMI	dep + PMI

表 8 第一组比较方法中，PCC-value[15] 和 STC-value[16] 为对 C-value 方法的改进。其中 PCC-value 方法将候选术语的文档频次融入 C-value 计算之中，以进行候选术语排序；STC-value 方法利用与候选术语有相同术语部件的相似候选术语信息，以提高 C-value 术语抽取效果。

表 8 第二组比较方法中，PMIK[19] 和 EMI[20] 为对 PMI 方法的改进。其中，PMI^k 方法使用联合概率因子，以改善 PMI 方法过高评估低频候选术语强度的问题；EMI 方法使用增强互信息 EMI 排序候选术语，改善互信息对称性不能很好衡量术语各成分内部的紧密程度问题。

实验结果如图 6、7 所示。由图 6 可见，pos2 + PCC-value 和 pos2 + STC-vlue 均略高于 pos2 + C-value，表明通过对 C-value 的改进，提高了专利术语抽取的准确性，例如，N = 1 000 时，分别比 pos2 + C-value 提高了 3.73% 和 4.83%； 而 dep + C-value 方法的准确率值最

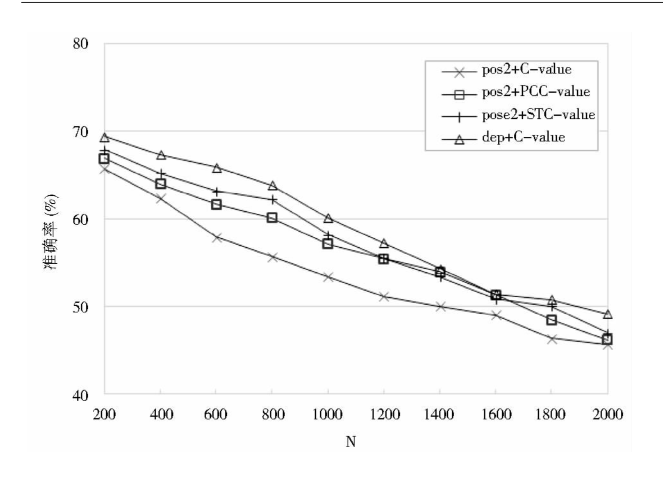


图6 第一组改进的候选术语选取和改进的候选术语排序方法比较

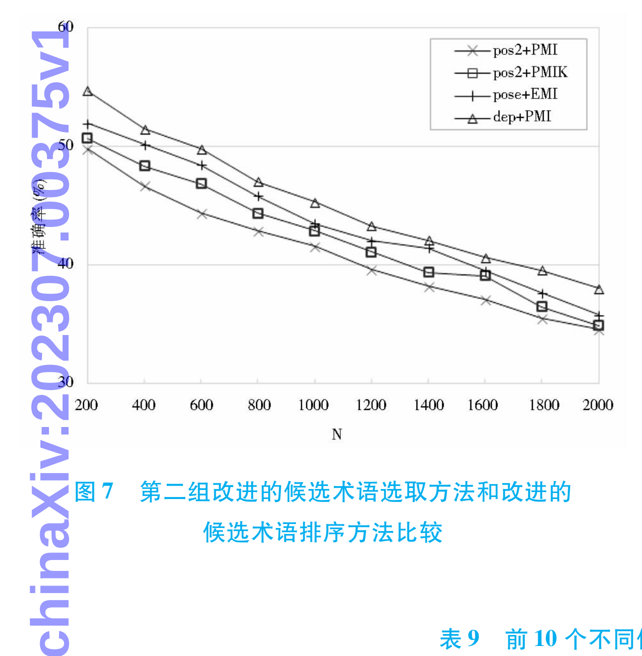


图7 第二组改进的候选术语选取方法和改进的候选术语排序方法比较

高，比 pos2 + C-value 提高了 6.77%。同样地，由图 7 可见，第二组实验得到了与第一组实验类似的结果，例如当 N = 1 000 时，pos2 + PMIk 和 pos2 + EMI 比 pos2 + PMI 分别提高了 1.35% 和 1.93%，而 dep + PMI 则比 pos2 + PMI 提高了 3.76%。两组实验表明，相对于目前研究主要针对专利术语抽取的第二个阶段，即候选术语排序方法的改进，第一个阶段的改进会对专利术语抽取产生较大的影响，从而影响后续专利术语排序，以及最终的专利术语抽取的准确性。因此，第一阶段的专利候选术语排序应该引起研究者的足够重视。

4.3.3 候选术语选取方法具体分析 最后，表 9 列出了 n-gram、NP、pos1、pos2 和 dep 选取的前 10 个最高频次的候选术语，其中正确的术语使用粗体表示。由表 9 可见，*n*-gram 去除一些不必要的词，使用词串的方式选取候选术语，造成较多的噪声数据；使用 NP 方法虽然能够过滤一些 *n*-gram 方法产生的噪声数据，但是仍会包含一些噪声，更为重要的是一些正确术语可能被漏选，如“氧化 石墨烯”，因为“氧化”为动词，而造成漏选；使用 pos1 和 pos2 的方法则在一定程度上克服了 NP 方法问题，包含了更加丰富的词性，但是这样在保证不漏选的前提下，又带来了额外的噪声数据，如“制备 石墨烯”因为其构成方式也是“动词 + 名词”而被选为候选术语；而本文提出的 dep 方法则能较好地克服这些问题，在不需要手工干预的前提下，选出更加准确的候选术语，克服动词等一些词的噪声干扰，为后续候选专利术语排序奠定坚实基础，最终获得更好的专利术语抽取效果。

表9 前10个不同候选术语选取方法选取

序号	n-gram	NP	pos1	pos2	dep
1	**氧化 石墨烯**	**离子 电池**	**氧化 石墨烯**	**氧化 石墨烯**	**氧化 石墨烯**
2	石墨烯 纳米	简单 工艺	制备 石墨烯	制备 石墨烯	**复合 材料**
3	石墨烯 复合	**锂离子 电池**	掺杂 石墨烯	掺杂 石墨烯	**改 性**
4	制备 石墨烯	**纳米 复合材料**	氮 掺杂	**纳米 复合材料**	**碳 纳米**
5	掺杂 石墨烯	优异 性能	**纳米 复合材料**	**复合 电极**	**三 维**
6	工艺 简单	**气相 沉积**	还原 氧化	制备 工艺	石墨烯 纳米
7	**超级 电容器**	**氟 乙烯**	**复合 电极**	**复合 薄膜**	**碳 纳米管**
8	石墨烯 量子	高 导电	纳米 颗粒	改性 石墨烯	**衬 底**
9	氮 掺杂	材料 领域	超声 分散	碳 纳米	**锂 离子**
10	石墨烯 分散	金属 基	制备 氧化	透明 导电	**离子 电池**

5 总结

依存句法分析通过语句单位内词语间的依存关系揭示词语间的语义修饰关系，进而实现对语义的理解，可以较为有效地弥补单纯依靠词性手段难以触及深层语义关系的不足。因此，本文首次将依存句法分析引入中文候选术语选取研究之中，提出基于依存句法分析的中文专利候选术语选取方法，以提高中文专利术语抽取的准确性。方法主要包括依存句法分析、剪枝、

生成依存子树等三个主要步骤。首先对中文专利进行依存句法分析,得到依存树,对依存树进行剪枝,去除不符合要求的依存关系,生成依存子树,从中选取连续词串作为候选术语,以抽取中文专利术语。本文提出的基于依存句法分析的方法能够找到较多正确的术语,且错误术语相对较少,从而为后续候选术语排序打下坚实基础。实验结果表明,本文提出的基于依存句法分析的中文候选术语选取方法相对而言,明显优于传统的候选术语选取方法,能够获得更高的中文专利术语抽取准确率。相对于目前研究主要针对专利术语抽取的第二个阶段,即候选术语排序方法的改进,第一个阶段的改进会对专利术语抽取产生较大的影响,从而影响后续专利术语排序,以及最终的专利术语抽取的准确性,应该引起研究者的足够重视。后续的研究将探索如何建立统一标准进行中文专利术语抽取正确性评估,以避免人工评估带来的主观性,更加客观地评估各种中文专利术语抽取方法的准确性。

参考文献:

[1] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International journal on digital libraries, 2000, 3(2):115－130.
[2] 周浪, 史树敏, 冯冲,等. 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010, 29(3):460－467.
[3] 韦小丽, 孙涌, 张书奎,等. 基于最大熵模型的本体概念获取方法[J]. 计算机工程, 2009, 35(24):114－116.
[4] 王昊, 王密平, 苏新宁. 面向本体学习的中文专利术语抽取研究[J]. 情报学报, 2016, 35(6):573－585.
[5] LI L, DANG Y, ZHANG J, et al. Domain term extraction based on conditional random fields combined with active learning strategy[J]. North American review, 2012, 174:368－375.
[6] CONRADO M, PARDO T, REZENDE S. A machine learning approach to automatic term extraction using a rich feature set[C]// The North American chapter of the Association for Computational Linguistics. Stroudsburg PA: Association for computational linguistics, 2013: 16－23.
[7] 胡阿沛, 张静, 刘俊丽. 基于改进 C-value 方法的中文术语抽取[J]. 现代图书情报技术, 2013, 29(2):24－29.
[8] 丁杰, 吕学强, 刘克会. 基于边界标记集的专利文献术语抽取方法[J]. 计算机工程与科学, 2015, 37(8):1591－1598.
[9] 刘剑, 唐慧丰, 刘伍颖. 一种基于统计技术的中文术语抽取方法[J]. 中国科技术语, 2014, 16(5):10－14..
[10] 曾镇, 吕学强, 李卓. 一种面向专利摘要的领域术语抽取方法[J]. 计算机应用与软件, 2016, 33(3):48－51.
[11] 杨双龙, 吕学强, 李卓,等. 中文专利文献术语自动识别研究[J]. 中文信息学报, 2016, 30(3):111－117.
[12] 徐川, 施水才, 房祥,等. 中文专利文献术语抽取[J]. 计算机工程与设计, 2013, 34(6):2175－2179.
[13] 张杰, 张海超, 翟东升. 面向中文专利权利要求书的分词方法研究[J]. 现代图书情报技术, 2014, 30(9):91－98.
[14] 胡文敏, 何婷婷, 张勇. 基于卡方检验的汉语术语抽取[J]. 计算机应用, 2007, 27(12): 3019－3020.
[15] 韩红旗, 朱东华, 汪雪锋. 专利技术术语的抽取方法[J]. 情报学报, 2011, 30(12):1280－1285.
[16] 俞琰, 赵乃瑄. 基于通用词与术语部件的专利术语抽取[J]. 情报学报, 2018, 37(7): 742－752.
[17] 林自芳, 蒋秀凤. 基于词内部模式的新词识别[J]. 计算机与现代化, 2010, 11(1):162－164.
[18] PECINA P, SCHLESINGER P. Combining association measures for collocation extraction. [C]// Proceedings of the COLING/ACL on main conference poster sessions. New York: ACM, 2016: 651－658.
[19] 杜丽萍, 李晓戈, 于根,等. 基于互信息改进算法的新词发现对中文分词系统改进[J]. 北京大学学报(自然科学版), 2016, 52(1):35－40.
[20] ZHANG W, YOSHIDA T, TANG X, et al. Improving effectiveness of mutual information for substantival multiword expression extraction[J]. Expert systems with applications an international journal, 2009, 36(8):10919－10930.
[21] ROBINSON J. Dependency structures and transformational rules [J]. Language, 1970, 46(2): 259－285.
[22] 白妙青, 郑家恒. 动词与动词搭配方法的研究[J]. 计算机工程与应用, 2004, 40(27): 70－72.
[23] 刘怀军,车万翔,刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007, 21(1): 79－84.
[24] 王慧泽, 龚声蓉, 刘纯平. 融合全局和局部的 Fisherfaces 方法[J]. 计算机工程与应用, 2008, 44(24): 194－196.
[25] CHE W, LI Z, LIU T. A Chinese language technology platform [C]//The 23th international conference on computational linguistics. New York: ACM, 2010: 3－16.
[26] AGARWAL B, PORIA S, MITTAL N, et al. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach[J]. Cognitive computation, 2015, 7(4):487－499.
[27] 冯冲, 廖纯, 刘至润,等. 基于词汇语义和句法依存的情感关键句识别[J]. 电子学报, 2016, 44(10): 2472－2476.
[28] 邓淑卿, 李玩伟, 徐健. 基于句法依存规则和词性特征的情感词识别研究[J]. 情报理论与实践,2018, 41(5): 137－142.
[29] QUAN C, WANG M, REN F. An unsupervised text mining method for relation extraction from biomedical literature[J]. Plos one, 2014, 9(7): 1－8.
[30] 李明耀, 杨静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程, 2016, 42(6): 201－207.
[31] 甘丽新, 万常选, 刘德喜,等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2): 284－302.
[32] 李超, 柴玉梅, 高明磊,等. 句法分析和深度神经网络在中文问答系统答案抽取中的研究[J]. 小型微型计算机系统, 2017

(6): 1341 - 1346.

[33] 刘雄, 张宇, 张伟男, 等. 基于依存句法分析的复合事实型问句分解方法[J]. 中文信息学报, 2017, 31(3): 140 - 146.

[34] KLEIN S, MCCONLOGUE K, SIMMONS R F. Co-occurrence and dependency logic for answering English questions[J]. Journal of the American Society for Information Science & Technology, 2014, 15(3):196 - 204.

[35] WANG J, ZHANG J, AN Y, et al. Biomedical event trigger detection by dependency-based word embedding[J]. Bmc medical genomics, 2016, 9(2):45 - 54.

[36] 高源, 席耀一, 李弼程. 基于依存句法分析与分类器融合的触发词抽取方法[J]. 计算机应用研究, 2016, 33(5): 1407 - 1410.

[37] 张仲华, 苏方方, 姬东鸿. 生物医学事件触发词识别研究[J]. 计算机应用研究, 2017, 34(3): 661 - 670.

[38] 张雷瀚, 吕学强, 李卓, 等. 领域本体术语的抽取方法研究[J]. 情报学报, 2014, 33(2): 167 - 174.

作者贡献说明:

俞琰:提出研究思路,设计研究方案,进行实验,撰写论文;

陈磊:收集数据、清洗数据;

姜金德:分析数据;

赵乃瑄:修改论文。

Research on the Selection of Chinese Patent Candidate Term Based on Dependency Syntax Parsing

Yu Yan[1,2] Chen lei[1] Jiang Jinde[3] Zhao Naixuan[1]

[1] Information Service Department, Nanjing Tech University, Nanjing 210009

[2] Computer Science Department, Chengxian College, Southeast University, Nanjing 211816

[3] School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract: [**Purpose/significance**] Aiming at the difficulties in making different pattern matching rules for different data sets and the low accuracy of Chinese patent term extraction, this paper proposes a selection method of Chinese patent candidate term based on dependency syntax parsing to improve the accuracy of Chinese patent term extraction. [**Method/process**] The method mainly includes three main steps: dependency syntax parsing, pruning and dependency subtree generation. Firstly, dependency syntax analysis was carried out on the Chinese patent text, from which dependency tree were obtained. Then, the dependency subtrees were generated by removing dependency relations which do not meet requirements. At last, the continuous word strings were selected as candidate terms to extract Chinese patent terms. [**Result/conclusion**] The experimental results show that compared with the existing related methods, the proposed method based on dependency syntax parsing can effectively improve the accuracy of Chinese patent term extraction.

Keywords: term extraction dependency syntax parsing Chinese patent candidate term selection